IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 4, Ver. II (Jul-Aug. 2014), PP 01-04
www.iosrjournals.org

User Search Goal Prediction based on Feedback Sessions

Sreerenjini. P. R[1], Prasanth. R. S [2]
[1] *PG Scholar,MCET*
[2] *Asst Prof,Department of CSE,MCET*

Abstract *: If the topic is very vast and the query is ambiguous, the users may be having different search goals when they submit the queries to a search engine. The goals of the users can be analysed in order to make the search engine results more relevant. In this paper, the Click through Logs are analyzed to create feedback Sessions. These feedback Sessions generated are used to create Pseudo documents . These Pseudo documents are further analysed and clustered based on Kmeans clustering algorithm. Then these clustered documents are checked for relevance evaluvation using Classified Average Precision(CAP).*

I. Introduction

Search engines try to match words used in queries with words found on pages or in links pointing to those pages when providing search results. Often, the order that pages are returned to a searcher are based upon an indexing of text on those pages, text in links pointing to those pages, and some measure of importance based upon link popularity.Before pages are served to a viewer, however, they may be reranked for one reason or another.Some possibilities may be Filtering of duplicate, or near duplicate, content ie,Search engines don't want the same page or content to fill search results, and pages that are substantially similar may be filtered out of search results. This is not technically a reranking of search results but is a type of filtering . Second is Removing multiple relevant pages from the same site It isn't uncommon for more than one page from a site to be relevant to a search query. Search engines try to limit the amount of pages displayed in search results from the same site. If there is more than one page from a site that ranks for a search, a search engine may show a second result from that site after the first result, indenting the second page, and inserting a link to "more results from this site."Additional results may not be shown.Based upon personal interests A search engine may try to rerank results for a search to a specific searcher based upon past searches and other tracked activity on the web from that person. This kind of reranking may rely upon a person logging on to a personalized search.The fourth type of reranking is based upon local interconnectivityThe search engine may grab results, and then reorder the top N (e.g., 100, 1000, etc.) search results based upon how they link between themselves.Sorting for country specific results It's possible that a searcher may wish to see results biased towards sites coming from a specific country. Someone could possibly explicitly choose a preference for a specific country, or the system may try to dynamically understand such a preference based upon IP address. The following patent application explores methods for reranking based upon country preferences.The sixth possibility is Sorting for language specific results Preferences regarding language may be set by the user in a browser, or through the search engine, or may be identified by the search engine while looking at the search query, the user interface, and characteristics of the search results.

The feedback session is analysed and the pseudo documents are generated from the titles and snippets of the Urls. Thus enriched Urls form the content of the pseudo documents.Based on the content of the pseudo documents they are clustered.The clustering of the search results is performed using k means clustering algorithm.

This paper is organized as follows. Section II contains various related works. Section III contains the procedures used in this paper. Proposed algorithm includes steps such as User Click Through Log Generation, Feedback session creation, Pseudo document generation, Similarity calculation ,Clustering and relevence calculation using CAP. Section IV discuss about the experiments and results obtained. Finally section V contains conclusions.

II. Related Works

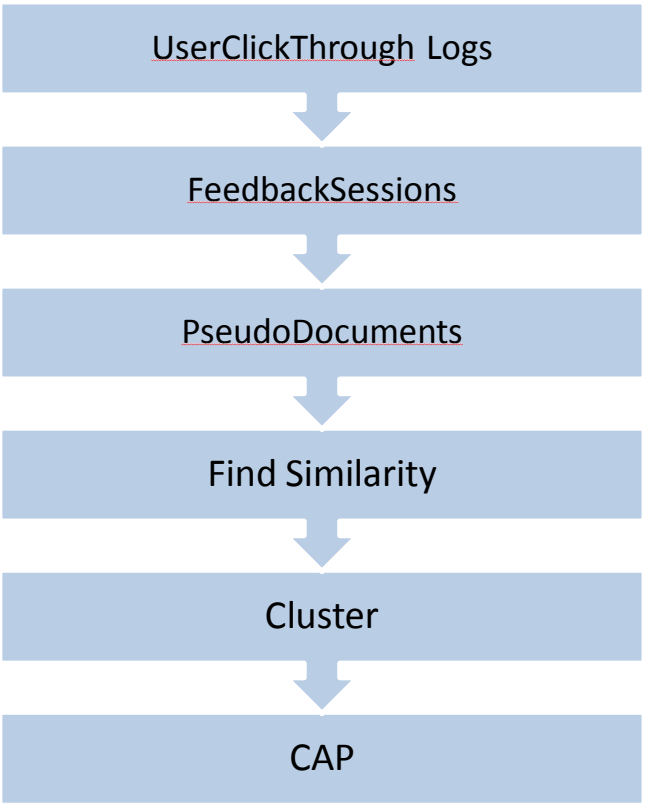
Oren Zamir,Oren Etzioni (1999) in their paper proposed Grouper, an interface to the results of the HuskySearch meta-search engine, which dynamically groups the search results into clusters labeled by phrases extracted from the snippets. In a method proposed by Ricardo Baeza-Yates,CarlosHurtado,Marcelo Mendoza[2005] , when a query is proposed so that when a query is submitted to a search engine, a list of related queries are suggested. These related queries are based on the previous issued queries. Groups of related queries found by running a clustering algorithm over the queries and their associated information in the logs. Hua-Jun Zeng,Qi-Cai He, Zheng Chen and Wei-Ying Ma[2004] states that Search result ranking problem is a salient phrase ranking problem.It converts an unsupervised clustering problem in to a supervised learning problem.A

study by Uichin Lee, Zhenyu Liu,Junghoo Cho[2005] of whether and how can we automate user goal identification process was surveyed.

III. The Proposed Algorithm

This section contains the proposed User Click Through Log Generation, Feedback session creation, Pseudo document generation, Similarity calculation ,Clustering and Relevence calculation using CAP. Figure.1 shows the flow chart of proposed algorithm. The interface allows only authorized users to enter. The users initially have to sign up into the interface and obtain a log account in order to utilize the facility. Once he has created his account he can access different links they require.

The User informations are captured and stored in the database.The user logs into the system and his access informations are also collected and stored in the database.These informations include Linkname,Login time,Log Out time,dwelling time, etc.These informations are utilized for the generation of feedback sessions. Users often leave Web pages in 10-20 seconds, but pages with a clear value proposition can hold people's attention for much longer because visit-durations follow a negative Weibull distribution.The dwell time is used for information retrieval tasks. Analysing the dwell time can be useful in improving the browsing experience. Several browsing features can be considered for improving search result relevance. In this paper both the number of clicks and dwell time on a web page are considered to rank the web page.



Figure, 1.Flow chart of proposed algorithm.

Users often leave Web pages in 10-20 seconds, but pages with a clear value proposition can hold people's attention for much longer because visit-durations follow a negative Weibull distribution.The dwell time is used for information retrieval tasks.Analysing the dwell time can be useful in improving the browsing experience.Several browsing features can be considered for improving search result relevance.In this paper both the number of clicks and dwell time on a web page are considered to rank the web page.

Clustering can be considered the most important unsupervised learning problem .It is the process of organizing objects into groups whose members are similar in some way A cluster is therefore a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters.

IV. Experimental Result

Input Dataset consists of access information and User information. Access information consists of 5 fields.They are Userid, password, date, time, linkname.User information consists of 10 fields.They are fno, first name, last name, contact no, pro, loc, country, email ,user id ,password.

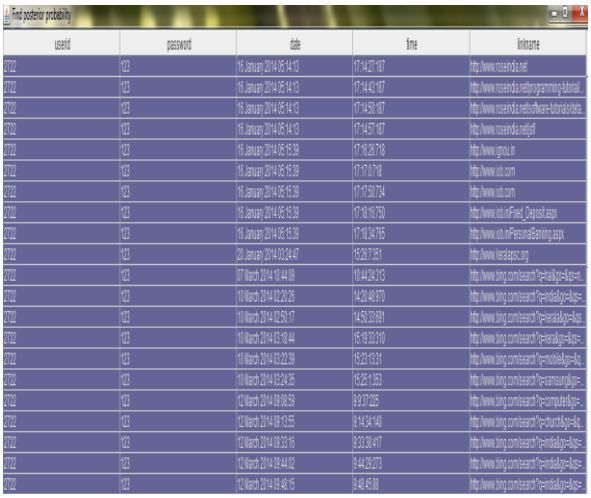


Figure.2. Information retrieved for Page relevance evaluation

Information retrieved during the user browsing session shown in figure.2 is utilized to create feedback session i.e., the webpage dwell time is used to decide the rank of the page. Using the titles and snippets in the web search results, the pseudo documents are created .Pseudo documents consists of the list of titles and snippets associated with each URL listed in the web search result page. These pseudo documents are then clustered using k means clustering algorithm.

Access No	Click Sequence
1	0
2	0
3	0
6	0
8	0
10	0
5	1
7	1
4	2

Figure.3 shows the feedback session for a particular browsing session.

User information retrieved during registration process can also be utilized (such as his 'profession') to cluster the page results. In each cluster the URLs can be arranged based on the ranks calculated based on the dwell time in the feedback sessions as shown in Figure.3.

Finally relevance for the pages are calculated and ranked accordingly. Figure 4 and Figure 5 shows the experimental results when k means clustering algorithm was applied on the web search results .The clustering is performed based on the similarity of the urls.

```
kMeans
======

Number of iterations: 2
Within cluster sum of squared errors: 0.0
Missing values globally replaced with mean/mode

Cluster centroids:
                                Cluster#
Attribute            Full Data          0          1          2
                         (170)       (50)      (110)      (10)
=============================================================================
group                   other       news      other entertainment



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       50 ( 29%)
1      110 ( 65%)
2       10 (  6%)
```

Figure.4 clustered search results using weka tool.

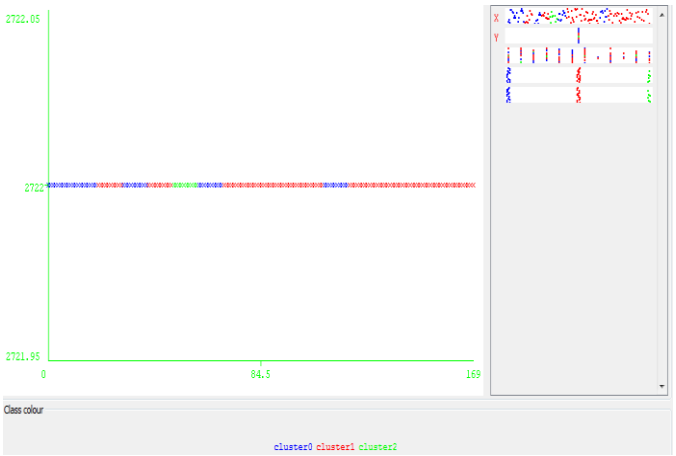


Figure 5. Visualisation of clustered results.

V. Conclusion

Both clicked and unclicked urls are considered as implicit feed backs in this paper .And the user information are also utilized to find the search result relevance. This system better analyses user search goals .Complexity is very low. Future extension involves the use of eye trackers to improve the search result relevance.

References

[1] Zheng Lu, Student Member, IEEE, Hongyuan Zha, Xiaokang Yang, Senior Member, IEEE,Weiyao Lin, Member, IEEE, and Zhaohui Zheng, " ,A New Algorithm for Inferring User Search Goals with Feedback Sessions", IEEE Transactions on IEEE Transactions on Knowledge and Data Engineering, vol 25, no. 3 ,March 2013.

[2] D. Beeferman and A. Berger," ,Agglomerative Clustering of a Search Engine Query Log", Proc. Sixth ACM SIGKDD Conf. Knowledge Discovery and Data Mining , 2000

[3] R. Baeza-Yates, C. Hurtado, and M. Mendoza ,S,"Query Recommendation Using Query Logs in Search Engines".

[4] Jaime Teevan MIT,Susan T. Dumais, (Microsoft Research,)Eric Horvitz(Microsoft Research)".Personalizing Search via Automated Analysis of Interests and Activities."

[5] Uichin Lee, University of California,Zhenyu Liu University of California ,Junghoo Cho University of California "Automatic Identification of User Goals in Web Search"

[6] J.-R Wen, J.-Y Nie, and H.-J Zhang, "Clustering User queries of a Search Engine," Proc. Tenth Int'l Conf. World Wide Web.

[7] Bing Pan,Helene A. Hembrooke,Geri K. Gay, Laura A. Granka, Matthew K. Feusner and Jill K. Newman, Information Science Program,"The Determinants of Web Page Viewing Behavior: An Eye-Tracking Study".

[8] X. Wang and C.-X Zhai, "Learn from Web Search Logs to Organize Search Results," Proc. 30th Ann. Int'l ACM SIGIR Conf.Research and Development in Information Retrieval (SIGIR '07),pp. 87-94, 2007.